SUPPLEMENTARY INFORMATION FOR:

High-density SNP association study and copy number variation analysis of the *AUTS1* and *AUTS5* loci implicates the *IMMP2L-DOCK4* gene region in autism susceptibility.

Maestrini E*, Pagnamenta AT*, Lamb JA*, Bacchelli E, Sykes NH, Sousa I, Toma C, Barnby G, Butler H, Winchester L, Scerri TS, Minopoli F, Reichert J, Cai G, Buxbaum JD, Korvatska O, Schellenberg GD, Dawson G, de Bildt A, Minderaa RB, Mulder EJ, Morris AP, Bailey AJ[§], Monaco AP[§], IMGSAC

* The first three authors contributed equally to this work

[§] to whom correspondence should be addressed.

E-mail: anthony.monaco@well.ox.ac.uk; anthony.bailey@psych.ox.ac.uk

This file includes:

- Members of the International Molecular Genetic Study of Autism Consortium (IMGSAC)
- Supplementary Methods
- Supplementary References
- Figures S1 to S4
- Tables S1 and S2

IMGSAC members:

Richard Holt[1], Kirsty Wing[1], Michael L Rutter[2], Jeremy R Parr[3], Simon Wallace[3], Kathy White[3], Marc Coutanche[3], Suzanne Foley[3], Katy Renshaw[3], Kerstin Wittemeyer[3], Magdalena Laskawiec[3], Patrick F Bolton[4], Gillian Baird[5], Vicky Slonims[5], Zoe Docherty[6], Stephen Abbs[6], Caroline Ogilvie[6], Pamela Warburton[6], Andrew Pickles[7], Jonathan Green[8], Catherine Aldred[8], Julie-Anne Wilkinson[8], Ann Le Couteur[9], Tom Berney[9], Helen McConachie[9], Emma Weisblatt[10], Lennart Pedersen[11], Demetrious Haracopos[11], Torben Isager[12], Birgitte Viskum[12], Ester Ulsted Sorensen[12], Karen Brondum-Nielsen[13], Bernadette Rogé[14], Jeanne Fremolle-Kruck[14], Carine Mantoulan[14], Inci Unsaldi-Cordier[14], Fritz Poustka[15], Gabriele Schmoetzer[15], Eftichia Duketis[15], Sven Boelte[15], Sabine Feineis Matthews[15], Sabine Schlitt[15], Annemarie Poustka[16], Sabine M Klauck[16], Bärbel Felder[16], Geeta Pakalapati[16], John Tsiantis[17], Katerina Papanikolaou[17], Elena Giouroukou[17], Elena Paliokosta[17], Simona Carone [18], Francesca Blasi[18], Fiorella Minopoli[18], Agatino Battaglia[19], Tiziana Filippi[19], Raffaella Tancredi[19], Barbara Parrini[19], Roberta Igliozzi[19], Herman Van Engeland[20], Maretha de Jonge[20], Chantal Kemner[20], Marjolijn Langemeijer[20], Channa Hijmans[20], Frederieke Koop[20], Wouter Staal[20], Catherine Lord[21], Edwin H Cook[22], Stephen J Guter[22], Jeff Salt[22], Bennett L Leventhal[22], Fred Volkmar[23], Eric Fombonne[24].

[1]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK; [2]Institute of Psychiatry, London, UK; [3]University Department of Psychiatry, Warneford Hospital, Oxford, UK; [4]Institute of Psychiatry, Department of Child and Adolescent Psychiatry, London, UK; [5]Guy's Hospital, Newcomen Centre, London, UK; [6]Regional Genetics Centre, Guy's Hospital, London, UK; [7]University of Manchester School of Epidemiology and Health Science, Manchester, UK; [8]Academic Department of Child Psychiatry, Booth Hall Children's Hospital, Manchester, UK; [9]University of Newcastle, Child and Adolescent Mental Health, Sir James Spence Institute, Newcaste upon Tyne, UK; [10]University of Cambridge Clinical School, Cambridge, UK; [11]Center for Autisme, Herlev, Denmark; [12]Børne- og ungdomspsykiatrisk center Glostrup, Denmark; [13]John F Kennedy Instituttet, Glostrup, Denmark; [14]Université de Toulouse Le Mirail, Centre d'Etudes et de Recherches en Psychopathologie (CERPP), Toulouse, France; [15]J W Goethe Universitaet Frankfurt, Klinik für Psychiatrie und Psychotherapie des Kindes- und Jugendalters, Frankfurt, Germany; [16]Deutsches Krebsforschungszentrum (DKFZ), Department of Molecular Genome Analysis, Heidelberg, Germany; [17]University Department of Child Psychiatry, Agia Sophia Children's Hospital, Athens, Greece; [18]Department of Biology, University of Bologna, Bologna, Italy; [19]Stella Maris Clinical Research Institute for Child and Adolescent Neuropsychiatry, Calambrone (Pisa), Italy; [20]Department of Child and Adolescent Psychiatry, University Medical Center, Utrecht, the Netherlands; [21]University of Michigan Autism and Communicative Disorders Center (UMACC), Ann Arbor, MI, USA; [22]University of Illinois at Chicago, Institute for Juvenile Research, Chicago, IL, USA; [23]Yale University Child Study Centre, New Haven, CT, USA; [24]McGill University Division of Psychiatry, Montreal Children's Hospital, Montreal, Quebec, Canada.

SUPPLEMENTARY METHODS

DNA samples

Genomic DNA was extracted from whole blood or EBV transformed lymphoblastoid cell lines. In a minority of cases DNA was extracted from buccal swabs. For 14% of DNA samples having low concentration (< 100ng/μl) whole genome amplification (WGA) was carried out using the GenomiPhi (GE Health Care) or Qiagen kit (REPLI-G). All samples were quantified using PicoGreen and subsequently normalised to 100 ng/μl.

Genotyping Quality Control

20 duplicate SNPs (10 for each chromosome) were genotyped in stages 1 and 2 to test for experiment-wise concordance10 chromosome X SNPs were also genotyped to estimate levels of mistyping. Two HapMap CEPH samples were added to each 96-well sample plate for quality control validation of inter-plate concordance and quality assurance. All SNPs were examined for genotyping quality. Pedstats[1] was used to calculate MAF, genotyping call rate and deviation from Hardy-Weinberg equilibrium. We excluded all SNPs with less than 90% genotyping rate, SNPs deviating from Hardy-Weinberg Equilibrium ($P < 0.001$) in the control population, and SNPs with more than 1 Mendelian error. DNA samples with less than 90% genotyping rate were excluded. All SNP cluster plots were visually inspected using the BeadStudio software, and SNPs with bad clustering or more than three clusters were excluded. SNPs with ambiguous clusters were sequenced using the dideoxy chain termination method in order to resolve Mendelian errors. In this way a number of SNPs were "rescued". Most SNPs with multiple clustering turned out to have a secondary SNP

nearby, in the allele-specific or locus-specific primer sequences; a few SNPs were triallelic.

In addition, genotyping success rates were evaluated in cases versus controls and in WGA versus non-WGA samples. No significant differences in missingness were detected.

Statistical Analysis

Power Calculations

Preliminary power calculations were performed in order to determine the optimum study design. Genotypes were simulated using an autism prevalence of 0.17% and allelic odds ratios (OR) ranging from 1.1 to 1.6. and a range of risk allele frequencies from 0.05 to 0.5.

Two alternative strategies were compared, given fixed genotyping resources totalling 1.8 million genotypes. *i*) Strategy A: type 200 controls and 100 sib-pair families (both affected sibs), chosen at random, at 3000 SNPs. *ii*) Strategy B: type 200 controls and 133 families (one affected sibs), selected for IBD, at 3000 SNPs. In 1000 replicates, strategy B ourperformed strategy A, for any combination of OR and risk allele frequencies. For instance, in the case-control analysis the proportion of 1000 replicates yielding significant evidence of association at 5% experimentwise significance level (assuming Bonferroni correction for 3000 tests) at an minor allele frequency of 0.2 and OR of 1.6 was 0.661 for IBD sharing families, but only 0.415 for randomly selected families. In the corresponding family-based calculations, the power was also increased for this alternative strategy; 0.225 compared to 0.099.

Power calculations for the IMGSAC-R and ND replication data sets were carried out using the Genetic Power Calculator for discrete trait TDT (GPC)[2]. Parameters used

for the GPC were 0.17% for the disease prevalence, a perfect LD between tested marker and disease allele, an additive model and a type 1 error rate of 0.0018 (applying the Bonferroni correction for 28 replication SNPs tested at each locus). Given that our significant SNPs in the primary sample showed an allelic OR of less than 2, and risk allele frequencies between 0.1 and 0.4, we performed power calculations over a range of risk allele frequencies (0.1-0.4) and of OR (1.2-2), allowing for the winners curse effect. This analysis showed that our combined replication sample (IMGSAC-R and ND collections) including 295 complete trios should give us enough power ($\geq$ 77%) to detect a risk allele with frequency > 0.2 and genotype relative risk (GRR) $\geq$ 1.7, or a risk allele with frequency > 0.1 and GRR $\geq$ 2.

While the Mount Sinai sample should provide sufficient power for replication due to its larger sample size (358 families), the 62 trios from U. Washington would allow only common variants with a large effect (GRR > 2) to be detected, although this might be an underestimation since these families were selected for increased IBD sharing on chromosome 7 from a larger collection of 222 families.

Comparison of LD levels between autism and HapMap CEU samples

To comprehensively evaluate how well the LD structure in the HapMap CEU data models the structure found in the IMGSAC autism sample, the level of LD (r^2) between all tag SNP pairs within 500kb of each other on chromosome 2 was calculated using HaploView and compared between the two populations. LD levels were significantly correlated between the two populations ($r^2 = 0.95$), indicating that the LD structure in the HapMap CEU data can be readily applied to our autism sample, although there may be a minor loss in capturing the variability.

GENEBPM software

The GENEBPM[3,4] algorithm was originally developed to assess haplotype association with disease in population-based studies within small candidate regions or blocks of strong LD. Maximum-likelihood SNP haplotype reconstructions are obtained via implementation of the expectation-maximisation (E-M) algorithm. A Bayesian partition model is used to describe the correlation between SNP haplotypes and causal variants at unobserved functional polymorphisms. Under this model, haplotypes are clustered according to their similarity in terms of marker-SNP allele matched, which is used as a proxy for shared ancestry. Haplotypes within the same cluster are assigned the same probability of carrying a causal variant. In this way, loss of power due to the presence of large numbers of haplotypes, particularly those that are rare, within a block can be avoided. Disease status is modelled in a logistic regression framework, parameterised in terms of additive and dominance effects of the causal variant(s), and here incorporating a main effect of gender. Evidence of association is assessed by means of a Bayes' factor, calculated by comparing the marginal likelihoods of a model of haplotype association (i.e. more than one cluster of haplotypes) with that of no haplotype association (i.e. one cluster of haplotypes). The methodology has also been extended to allow for family-based association studies for samples of trios. The SNP haplotypes in founders are reconstructed using an E-M algorithm, conditioning on the genotype data available in their child. Pseudo-control individuals are matched to each affected child, constructed from the possible haplotype pairs not transmitted to the affected child. Disease status is then modelled in a conditional logistic regression framework, parameterized in terms of additive and dominance effects of the causal variants, as before, but also in terms of parent of

origin effects, to allow for differential transmission of causal variants from the mother and father.

GENEBPM analyses were also performed using a sliding window of 5 SNPs across each chromosomal region. For comparison with frequentist single-SNP analyses, the GENEBPM algorithm has also been applied to each SNP in turn (i.e. single SNP "haplotypes").

Copy number variation

QuantiSNP analysis

After exclusion of whole genome amplified samples from the BeadStudio project (approximately 14% of samples), final reports were generated containing B-allele frequency, log R ratio data and build 36 genome coordinates. Data from both GoldenGate arrays were combined for each region, no-calls were deleted and these files were run on QuantiSNP v1.[5] using the following settings: L=1M, array type=100k, EMiters=25, maxcopy=4, GC correction=ON. The QuantiSNP software is not designed specifically for the GoldenGate platform and indeed, the number of CNVs detected suggested a high level of false positives. Therefore DNA samples resulting in a number of CNV $\geq$ 95th percentile ($\geq$5 CNVs per sample) were removed. CNVs were then sorted by log Bayes factor and any scoring less than 10 were removed from further analysis.

CNV validation and screening

Screening of the deletion identified in *UPP2* was carried out by multiplex PCR with two primer pairs: one inside the deleted region giving a fragment of 236 bp in the wild type allele, and the other across the deleted region giving rise to a fragment of 323 bp

in the deleted allele (Fig S2B).

Validation of CNVs in the *IMMP2L-DOCK4* region was performed by Quantitative Multiplex PCR of Short fluorescent Fragments (QMPSF)[6]. Short fragments in exons 2, 3, and 6 of the *IMMP2L*, exon 4 of *LRRN3* and exon 52 of *DOCK4* were simultaneously PCR amplified, in a single tube, using dye-labelled primers. An additional fragment, corresponding to exon 7 of the *RNF20* gene located on chromosome 9q, was co-amplified as a control. Sequences and PCR conditions of all primer pairs are available on request from the authors. One μl of the PCR product was resuspended in a mix containing 8.8 μl of deionised formamide, 0.2 μl of GeneScan™ 600 LIZ Size Standard (Applied Biosystems). PCR products were run on an ABI prism 3730 sequencer and the data analysed using GENEMAPPER™ software (Applied Biosystems).

The copy-number was calculated using the following formula:

(A sample/A average)x(A_{RNF} average/A_{RNF} sample), where "A sample" is the peak area of each locus-specific probe, while "A average" is the average of peak areas of all samples in the same run for the same probe. A reduction greater than 0.4 of the peak area indicates a hemizygous deletion, whereas a duplication results in an increase of at least 1.4. Each positive result was confirmed in a second independent QMPSF assay.

The segregation of the *IMMP2L-DOCK4* deletion in pedigree 15-0084 was also consistent with the inheritance pattern of two SNPs (rs1978247 and rs12672270).

Quantitative PCR of *DOCK4* exons 37, 31, 14 and 7 was carried out in blood-derived DNA from subject 15-0084-001 (father), 15-0084-002 (mother) and 15-0084-003 (affected son). The following primers were designed for various *DOCK4* exons using Primer3. Exon7, TGGGTCCTGTTATTTCCTTCAG and

TTCATCTGGACAAAGAGGTGGT; Exon14, AACCTGTGTGTTCTTCCCTTTG and GACCACCTGGGACTGTTGTTAT; Exon31, CTCACTTTAGGAGAGCACAAGC and TCTGCTCCCAGTCCATCATATC; Exon37, ATGACGAGCTACTGGAATGGTC and CCTCTGTCAAAGTTCTGGATGA. The housekeeping gene *GAPDH* was also amplified as a reference gene, using primers TACTAGCGGTTTTACGGGCG and TCGAACAGGAGGAGCAGAGAGCGA. Annealing temperature was 59°C for all PCRs, and efficiencies were calculated using a dilution series of 100ng, 50ng, 25ng, 12.5ng and 6.25ng and 3.125ng of DNA template. All samples were run in triplicate on the iQ5 Real-Time PCR Detection System (BioRad) with the cycle threshold (Ct) means used for calculations. We used the iQ SYBR Green Supermix, and carried out melting curve analysis of PCR products to ensure specific amplification. Ct outliers were removed if the SD of the triplicates was >0.5 cycles. Relative copy number was calculated taking PCR efficiency into account, and using the father (15-0084-001) as a non-deleted reference sample, as described[7]. The experiment was carried out twice and identical results were obtained.

Mutation screening

The entire coding sequence and putative regulatory regions of the *NOSTRIN* and *ZNF533* genes were sequenced in 31 autistic individuals, including all individuals carrying 2 copies of most significant risk alleles; the coding sequence of *UPP2* was sequenced in 47 subjects, including 12 probands carrying the deletion of exons 6 and 7; the *IMMP2L* and *LRRN3* genes were sequenced in eight individuals with autism, including five homozygous for the most significant risk haplotype. No novel coding variants were identified, except one silent change in exon 4 of *UPP2* gene in only one

individual.

Sequencing primers and conditions are available on request.

SUPPLEMENTARY REFERENCES

1. Wigginton JE,Abecasis GR. PEDSTATS: descriptive statistics, graphics and quality assessment for gene mapping data. *Bioinformatics* 2005; **21**: 3445-7.
2. Purcell S, Cherny SS, Sham PC. Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. Bioinformatics. 2003; **19**: 149-50.
3. Morris AP. Direct analysis of unphased SNP genotype data in population-based association studies via Bayesian partition modelling of haplotypes. *Genet Epidemiol* 2005; **29**: 91-107.
4. Morris AP. A flexible Bayesian framework for modeling haplotype association with disease, allowing for dominance effects of the underlying causative variants. *Am J Hum Genet* 2006; **79**: 679-94.
5. Colella S, Yau C, Taylor JM, Mirza G, Butler H, Clouston P *et al.* QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res* 2007; **35**: 2013-25.
6. Saugier-Veber P, Goldenberg A, Drouin-Garraud V, de La Rochebrochard C, Layet V, Drouot N *et al.* Simple detection of genomic microdeletions and microduplications using QMPSF in patients with idiopathic mental retardation. *Eur J Hum Genet* 2006; **14**: 1009-17.
7. Pfaffl MW. A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res* 2001; **29**: e45.

Supplementary Figure Legends

Figure S1: Graphical representation of chromosome 2 and 7 GENEPBM analysis

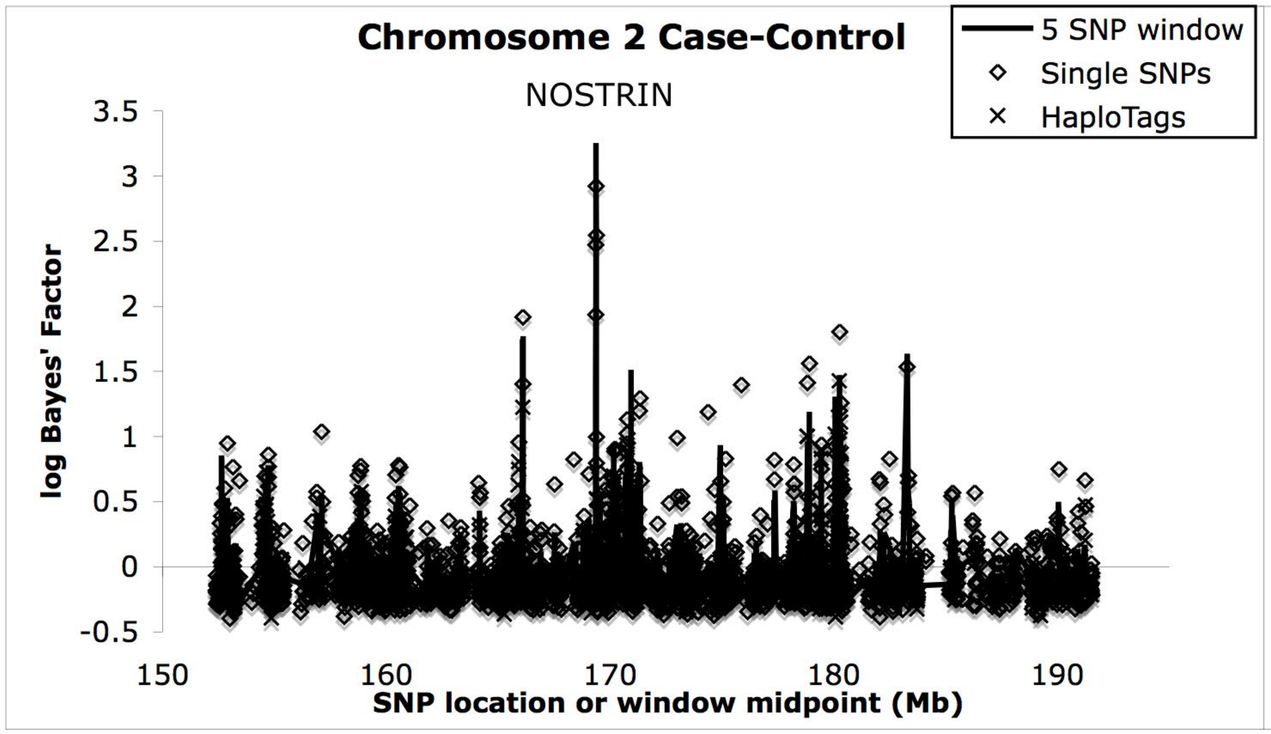
$\log_{10}$ Bayes' Factor values are plotted against the chromosome position.

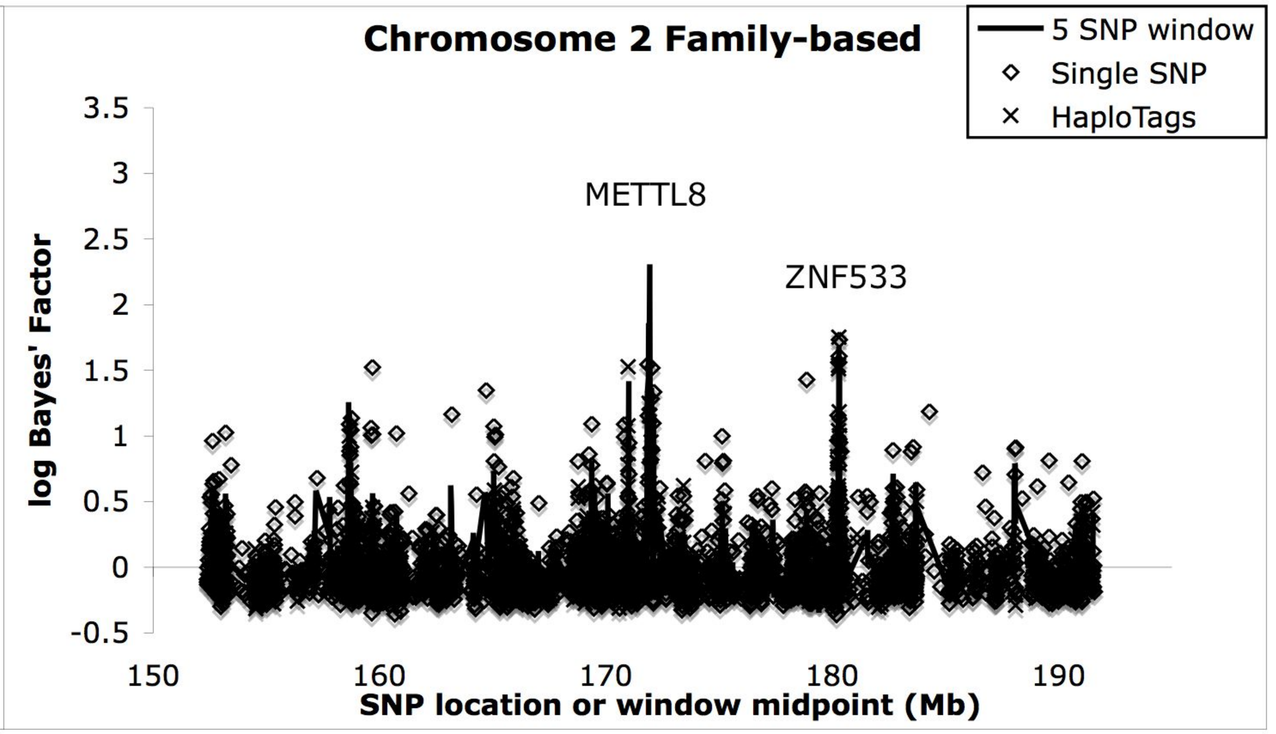
Figure S2: deletion in *UPP2*

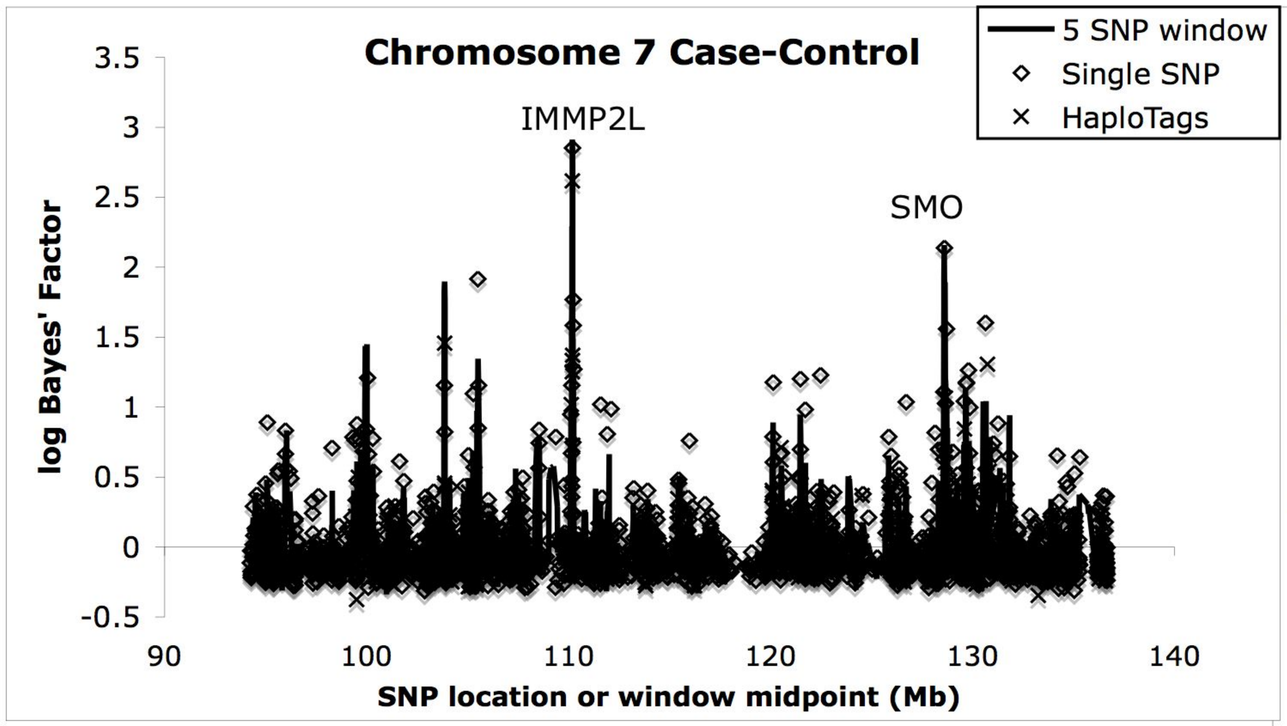
A: Position of the *UPP2* deletion on chromosome 2, shown on the UCSC browser; B: multiplex PCR assay showing 3 samples heterozygous for the deleted allele (lanes 4, 5 and 8)

Figure S3: SNP data showing *IMMP2L* / *DOCK4* duplication in subject 13-3023-001. GoldenGate SNP data for the proband is shown on both arrays separately. Data for SNPs showing B-allele frequency consistent with AAB or ABB genotypes are boxed. SNPs within the boundaries of the region detected are highlighted in red. Screenshots are from the Illumina Genome Viewer in BeadStudio.

Figure S4: Quantitative PCR of *DOCK4* exons in pedigree 15-0084. Quantitative PCR of blood-derived DNA from family 15-0084 indicates that the distal *DOCK4* deletion breakpoint is between exon 31 and exon 14. A relative copy number of 0.7 was used as the threshold for determining deleted regions. 15-0084-001 was used as the non-deleted reference sample.







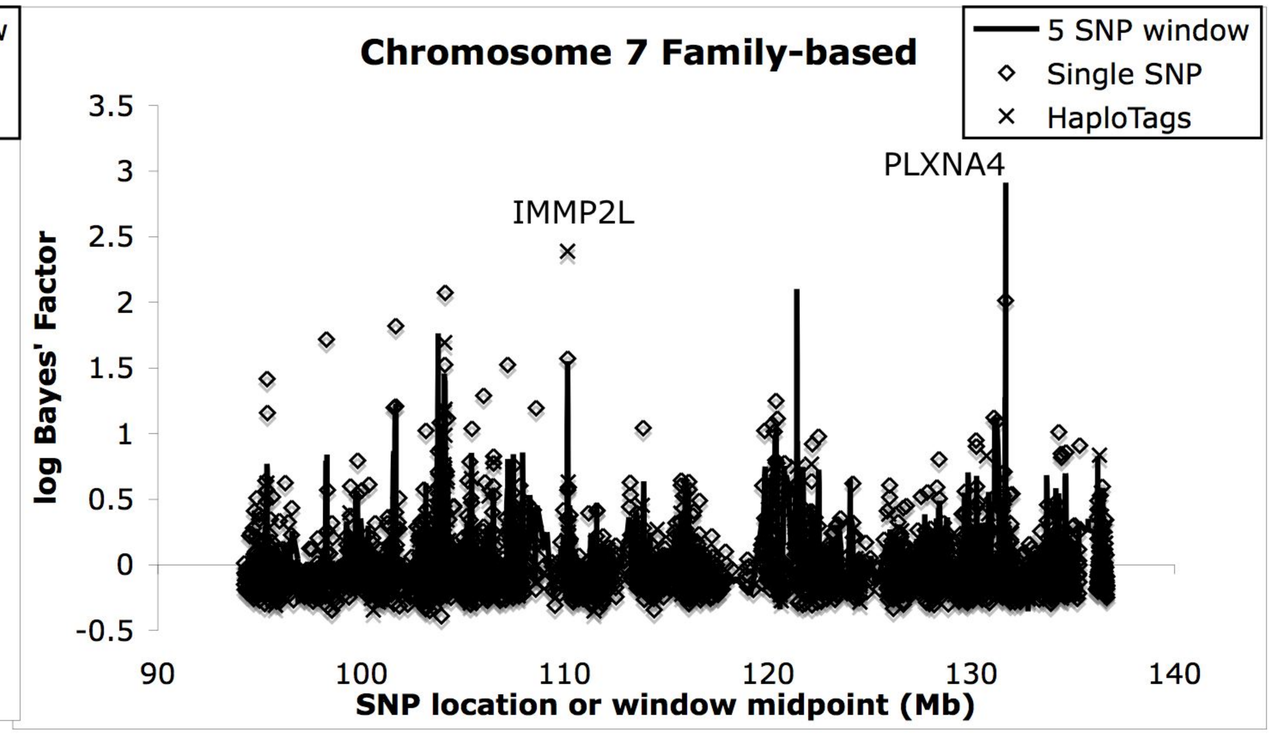
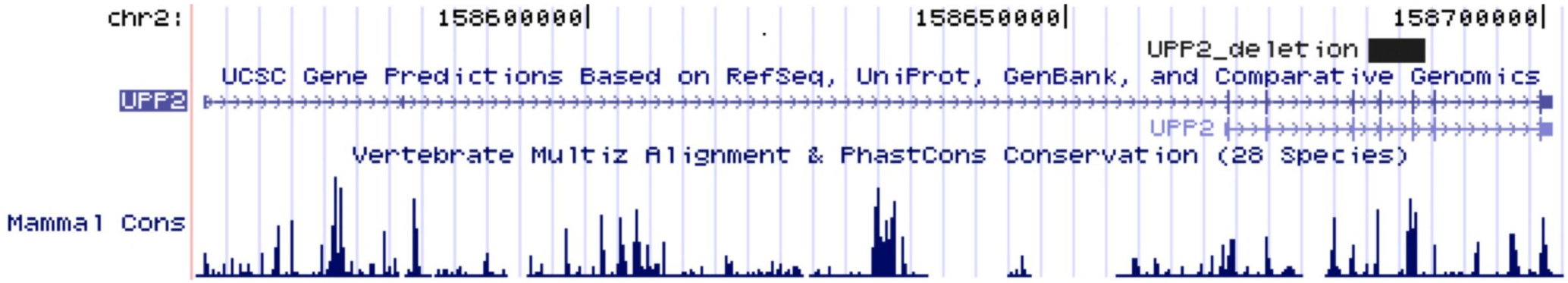


Figure S1: GENEPBM analysis

A



B

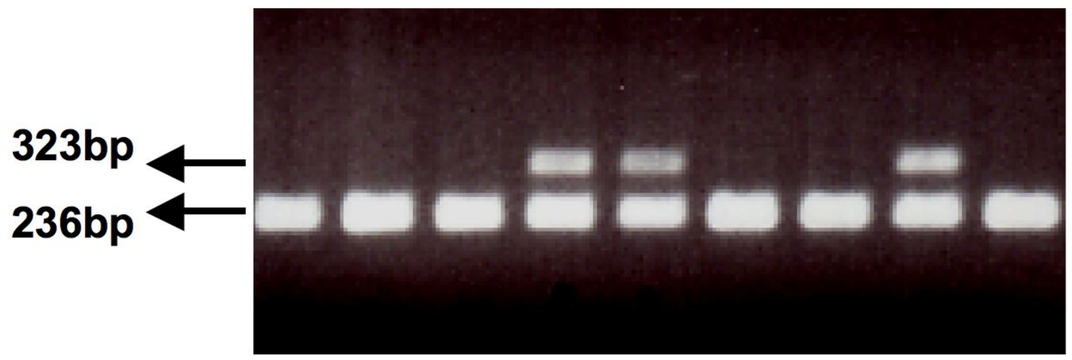


Figure S2: deletion in *UPP2*

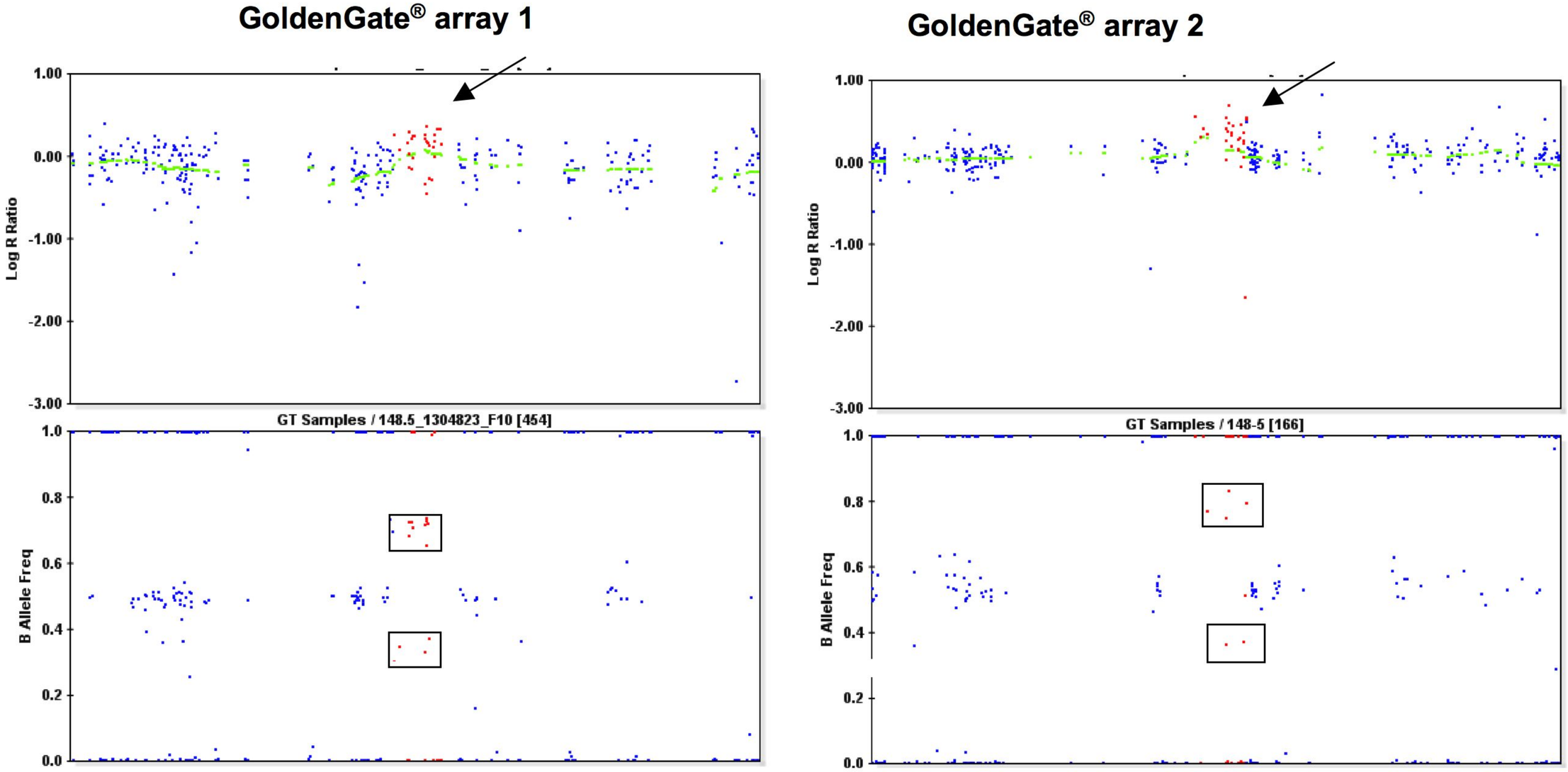


Figure S3: SNP data showing *IMMP2L* / *DOCK4* duplication in subject 13-3023-001.

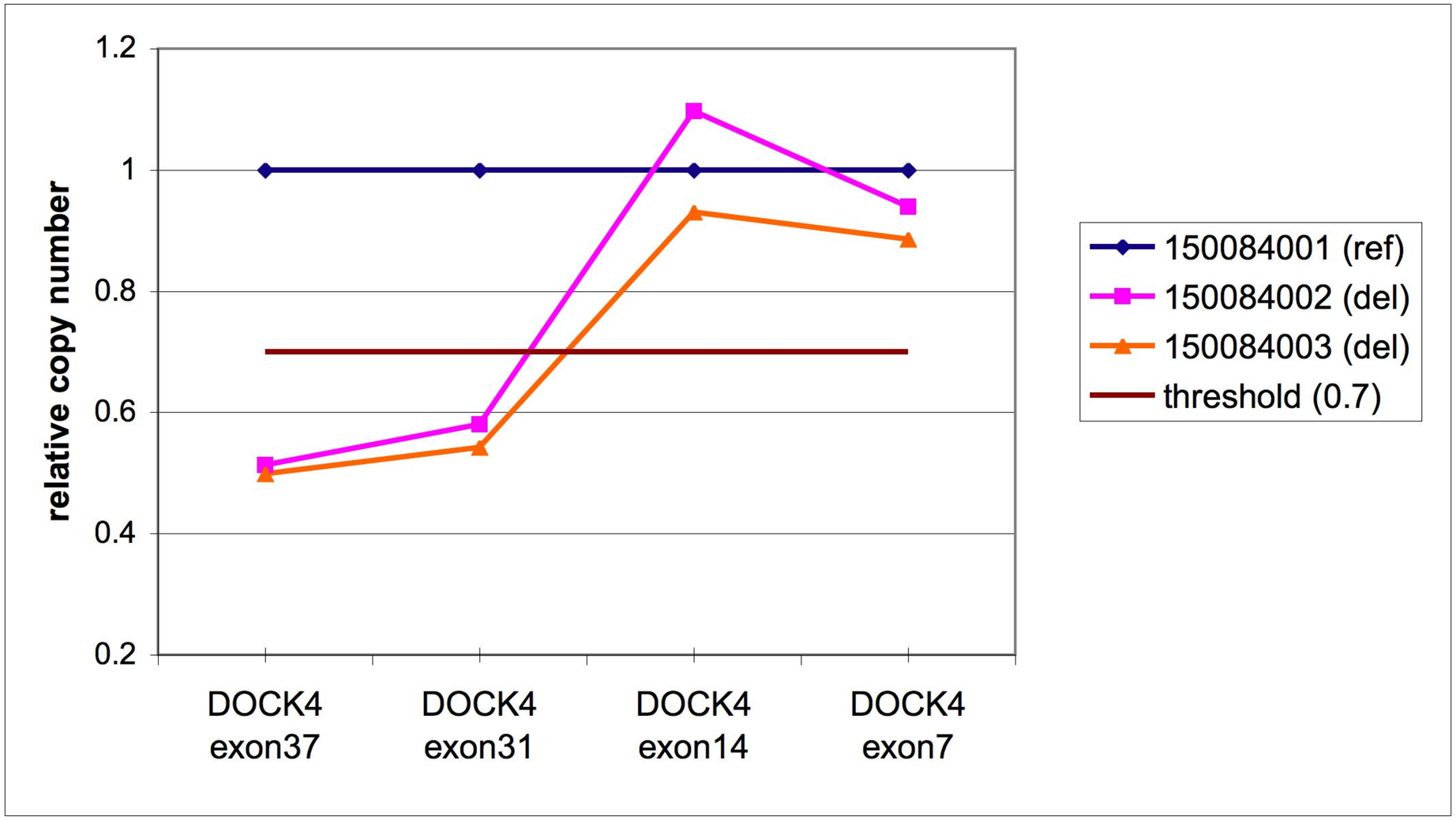


Figure S4. ***DOCK4*** **copy number determined by qPCR**

Table S1. Results of family-based analysis of replication samples using UNPHASED.

The table shows the nominal *P*-value for the likelihood ratio statistic, the allele frequency in affected offsprings and in untransmitted parental alleles (Ca-Freq, Co-Freq). Frequencies are reported for the risk allele detected in the primary association analysis. Flip-flop of associated allele is flagged by an asterisk. Nominal *P*-values<0.05 are in bold. UNPHASED family based analysis of the primary IMGSAC sample is in italics.

					IMGSAC-PRIMARY *(family-based analysis)*			***IMGSAC-R*** *(294 affected subjects)*			***ND*** *(96 affected subjects)*			***IMGSAC-R + ND*** *(390 affected subjects)*		
SNP	**Chr.**	**Gene**	**Alleles**	**Risk Allele**	***P-value***	***Ca-Freq***	***Co-Freq***	***P*-value**	**Ca-Freq**	**Co-Freq**	***P*-value**	**Ca-Freq**	**Co_Freq**	***P*-value**	**Ca-Freq**	**Co_Freq**
rs4664599	2	FMNL2	C/T	T	*0.0094*	*0.730*	*0.613*	0.2014	0.745	0.705	0.1891	0.698	0.635	0.0747	0.732	0.689
rs1427395	2	phastCons[c]	A/T	T	*0.0066*	*0.540*	*0.416*	0.3634	0.564	0.532	**0.0216**	0.500	0.387	0.0505	0.547	0.496
rs6437133	2	UPP2	C/T	C	*0.0046*	*0.632*	*0.504*	0.1630	0.543	0.500	**0.0395**	0.482*	0.599*	0.7106	0.529	0.519
rs6709528	2	UPP2	C/T	T	*7.84E-04*	*0.603*	*0.451*	0.3708	0.498	0.471	0.0902	0.458	0.545	0.9158	0.487	0.490
rs12620556	2	LOC130940	A/G	A	*0.0108*	*0.952*	*0.889*	0.8454	0.899	0.905	**0.0235**	0.905*	0.965*	0.2247	0.901	0.920
rs764660	2	SCN2A	A/C	C	*0.0050*	*0.373*	*0.255*	0.6841	0.326	0.334	0.8013	0.303	0.290	0.8341	0.320	0.325
rs1020626	2	FAM130A2	C/T	T	*0.2479*	*0.508*	*0.455*	0.2057	0.437	0.391	0.5230	0.365	0.397	0.3869	0.419	0.396
rs10930170	2	FAM130A2	A/G	G	*0.4706*	*0.397*	*0.365*	0.3667	0.330	0.299	0.8628	0.307	0.299	0.3943	0.324	0.302
rs829958 (proxy for rs829957)	2	NOSTRIN	A/G	G	*0.0137*	*0.220*	*0.130*	0.5600	0.140	0.129	0.2149	0.174	0.127	0.2435	0.149	0.127
rs6433093	2	NOSTRIN	A/C	A	*0.1133*	*0.344*	*0.276*	0.5168	0.298	0.276	0.2759	0.300	0.249	0.2621	0.299	0.266
rs7583629	2	NOSTRIN	A/G	A	*0.0027*	*0.262*	*0.148*	0.4710	0.172	0.157	0.7947	0.172	0.162	0.4405	0.172	0.155
rs482435[a]	2	NOSTRIN	C/T	C	*0.0156*	*0.381*	*0.278*	0.3782	0.333	0.302	0.9429	0.309	0.305	0.3995	0.327	0.301
rs12692976	2	SLC25A12	G/T	T	*0.0106*	*0.675*	*0.562*	0.6697	0.647	0.657	0.5861	0.601	0.572	0.9522	0.635	0.637
rs17705978	2	RAPGEF4	A/G	G	*0.2701*	*0.070*	*0.075*	0.6958	0.070	0.075	0.0912	0.026	0.062	0.2862	0.058	0.072
rs6758260	2	RAPGEF4	A/G	A	*0.0142*	*0.189*	*0.179*	0.7307	0.189	0.179	0.5899	0.153	0.133	0.5662	0.179	0.166
rs13007575	2	phastCons[c]	A/G	A	*0.0034*	*0.960*	*0.890*	0.1337	0.921	0.946	**0.0063**	0.958	0.886	0.8726	0.931	0.929
rs1866925 (proxy for rs6717587)	2	phastCons[c]	G/T	G	*0.0376*	*0.885*	*0.820*	0.3546	0.790	0.768	0.5508	0.821	0.798	0.2756	0.799	0.774
rs1434087[a]	2	OSBPL6	C/T	T	*0.0679*	*0.960*	*0.923*	**0.0399**	0.928	0.890	0.7597	0.916	0.925	0.0988	0.925	0.898
rs2276573[a]	2	ZNF533	G/T	T	*0.0184*	*0.631*	*0.531*	0.1910	0.585	0.545	0.1021	0.516	0.605	0.7569	0.569	0.561
rs7590028[a]	2	ZNF533	C/T	T	*0.0688*	*0.579*	*0.500*	0.2749	0.519	0.482	0.1572	0.464	0.541	0.8250	0.505	0.496
rs11885327	2	ZNF533	C/T	C	*5.53E-04*	*0.722*	*0.574*	0.4270	0.629	0.597	0.9218	0.621	0.626	0.5295	0.627	0.607
rs1964081[a]	2	ZNF533	A/G	A	*8.43E-04*	*0.901*	*0.791*	0.7606	0.872	0.870	0.4805	0.853	0.826	0.8637	0.867	0.859
rs11674376	2	ZNF533	C/T	C	*0.0029*	*0.385*	*0.258*	0.9504	0.277	0.275	0.9965	0.224	0.224	0.9528	0.263	0.263
rs723390	2	nn (Affy 10K)	A/G	A	*0.0073*	*0.628*	*0.498*	0.8900	0.568	0.575	0.4667	0.537	0.574	0.6395	0.560	0.572
rs3755248	2	TFPI	C/T	T	*0.0039*	*0.718*	*0.597*	0.2140	0.668	0.636	0.7806	0.626	0.613	0.1937	0.658	0.628
rs16829088 (proxy for rs7573488)	2	TFPI	A/G	G	*0.0034*	*0.794*	*0.684*	0.9973	0.732	0.730	0.2738	0.702	0.647	0.5838	0.724	0.710
rs3811608	2	FLJ20160	C/T	T	*0.0045*	*0.877*	*0.782*	0.0712	0.820	0.773	nd	nd	nd	nd	nd	nd
rs6757698	2	FLJ20160	C/T	C	*0.0065*	*0.369*	*0.264*	0.7289	0.288	0.304	0.3804	0.358	0.314	0.9484	0.305	0.305
rs12538145	7	SLC25A13	C/G	C	*0.0036*	*0.988*	*0.946*	0.3958	0.967	0.957	0.5059	0.937	0.953	0.7627	0.959	0.956
rs2307355	7	MCM7	A/G	A	*0.2060*	*0.083*	*0.054*	0.3797	0.031	0.041	0.6476	0.047	0.057	0.3223	0.035	0.046
rs11768465	7	FBX024	C/T	C	*0.0269*	*0.840*	*0.763*	0.3915	0.785	0.765	**0.0184**	0.761*	0.863*	0.6561	0.779	0.790
rs875659[b]	7	CUX1	C/T	C	*1.23E-04*	*0.524*	*0.355*	0.4509	0.445	0.483	0.9677	0.469	0.471	0.4977	0.451	0.478
rs3819479	7	RELN	A/T	T	*0.0023*	*0.821*	*0.703*	0.1745	0.746	0.782	0.9647	0.795	0.793	0.2366	0.759	0.788
rs6976167	7	LHFPL3	C/T	T	*0.3846*	*0.140*	*0.114*	0.4676	0.105	0.085	0.1805	0.109	0.069	0.2009	0.106	0.082
rs12666599	7	LHFPL3	C/T	T	*0.0030*	*0.901*	*0.804*	0.9622	0.846	0.847	0.4967	0.862	0.836	0.6922	0.850	0.843
rs4730037	7	LHFPL3	C/T	C	*0.0042*	*0.448*	*0.314*	0.2501	0.367	0.336	0.8526	0.344	0.354	0.3516	0.362	0.340
rs9690688	7	LAMB4	G/T	T	*0.0038*	*0.099*	*0.033*	0.3183	0.045	0.058	0.8242	0.068	0.074	0.3409	0.051	0.063
rs6951925	7	NT_007933.689[d]	G/T	G	*0.0027*	*0.242*	*0.134*	0.6671	0.152	0.169	0.2490	0.250	0.197	0.9084	0.175	0.173
rs1464895	7	IMMP2L	A/G	A	*0.1754*	*0.202*	*0.154*	0.4854	0.161	0.145	**0.0042**	0.120*	0.235*	0.3043	0.150	0.170
rs10499993[b]	7	IMMP2L	A/G	G	*0.3274*	*0.806*	*0.768*	0.0700	0.791	0.743	0.8217	0.755	0.766	0.1376	0.782	0.749

Table S1. Results of family-based analysis of replication samples using UNPHASED (continued).

SNP	Chr.	Gene	Alleles	Risk Allele	*IMGSAC-PRIMARY* (family-based analysis)			*IMGSAC-R* (294 affected subjects)			*ND* (96 affected subjects)			*IMGSAC-R + ND* (390 affected subjects)		
					P-value	***Ca-Freq***	***Co-Freq***	***P*-value**	**Ca-Freq**	**Co-Freq**	***P*-value**	**Ca-Freq**	**Co_Freq**	***P*-value**	**Ca-Freq**	**Co_Freq**
rs2030781[b]	7	IMMP2L	C/T	C	*0.0759*	*0.282*	*0.209*	0.0798	0.261	0.215	0.1661	0.184	0.243	0.3332	0.242	0.222
rs12537269	7	IMMP2L	A/G	A	*0.0482*	*0.294*	*0.213*	**0.0485**	0.262	0.210	0.7737	0.255	0.243	0.0667	0.260	0.220
rs1528039	7	IMMP2L	C/T	C	*0.1474*	*0.250*	*0.193*	0.3420	0.202	0.182	0.3065	0.201	0.247	0.8061	0.202	0.199
rs2217262	7	DOCK4	A/C	A	*0.0184*	*0.972*	*0.925*	**0.0272**	0.955	0.924	**0.0055**	0.979	0.916	**9.21E-04**	0.962	0.921
rs989613	7	NT_007933.632[d]	A/G	G	*0.0152*	*0.832*	*0.740*	0.2219	0.838	0.863	0.2199	0.739	0.795	0.0925	0.812	0.848
rs7807053	7	KCND2	A/C	A	*0.0611*	*0.107*	*0.062*	0.0677	0.094	0.132	0.5692	0.113	0.095	0.2002	0.099	0.122
rs41620	7	3' of TSPAN12	A/G	A	*0.0012*	*0.790*	*0.661*	0.7367	0.778	0.773	0.4398	0.714	0.747	0.9775	0.762	0.766
rs2525720	7	ING3	A/G	A	*0.0011*	*0.849*	*0.729*	0.3414	0.838	0.818	nd	nd	nd	nd	nd	nd
rs538558	7	3' of FEZF1	A/T	A	*0.1223*	*0.373*	*0.304*	0.8246	0.364	0.369	0.9714	0.319	0.317	0.9373	0.353	0.352
rs11978485	7	3' of SLC13A1	A/G	G	*0.0136*	*0.865*	*0.780*	0.6299	0.825	0.813	0.4272	0.771	0.733	0.3158	0.812	0.791
rs6962740	7	5' of SMO	C/G	G	*0.2366*	*0.849*	*0.812*	0.8941	0.791	0.783	0.2092	0.721	0.780	0.5698	0.772	0.783
rs2030974	7	5' of CPA2	C/T	C	*0.0150*	*0.401*	*0.294*	0.0515	0.279	0.333	0.9839	0.258	0.259	0.0991	0.274	0.314
rs2171493	7	5' of CPA2	A/C	C	*0.0233*	*0.337*	*0.243*	**0.0230**	0.242*	0.301*	0.8506	0.216	0.224	**0.0459**	0.235*	0.282*
rs10275276 (proxy for rs1863009)	7	AK054623	C/T	T	*0.2947*	*0.274*	*0.232*	0.9594	0.253	0.252	0.3956	0.247	0.211	0.6972	0.251	0.243
rs7787173	7	NT_007933.1017[d]	A/T	A	*0.0020*	*0.913*	*0.818*	0.3798	0.885	0.867	nd	nd	nd	nd	nd	nd
rs4731863	7	PLXNA4	A/T	T	*1.30E-04*	*0.976*	*0.895*	0.1591	0.907	0.931	0.0987	0.891	0.938	**0.0391**	0.903*	0.934*

[a] SNPs also tested in Mount Sinai sample (not significant)

[b] SNPs also tested in U.Washington sample (not significant)

[c] PhastCons - highly conserved region

[d] Transcript not annotated in human reference sequence build 36.1

Table S2. CNVs detected by QuantiSNP analysis

Chr.	Locus	Start (Mb)	End (Mb)	Length (kb)	log BF	Copy number	Sample ID		Segregation [a]	Gene(s)	DGV (overlap)
2	1	154.645	154.798	152.4	29.5	3	133010002	Mother	Not transmitted to proband	*GALNT13*	Yes (14%)
	2	168.721	169.384	663.5	183.9	3	C0953	Control	N/A	*STK39, LASS6*	Yes (24%)
	3	179.106	179.381	275.1	13.8	3	150033002	Mother	Not transmitted to proband	*TTN*	Yes (72%)
		179.250	179.381	130.8	11.3	3	130540004	Mother	Not transmitted to proband	*TTN*	Yes (100%)
	4	190.853	190.944	91.0	20.6	3	150022001	Father	Not transmitted to proband	*HIBCH, INPP1*	No
	5	191.224	191.538	313.9	10.4	1	130079004	Father	Not transmitted to proband	*NAB1, GLS*	Yes (2%)
7	1	100.749	100.917	167.7	48.0	3	140713301	Proband	Affected sib inherited duplication	*EMID2* and *RABL5*	Yes (100%)
		100.749	100.896	147.1	13.8	3	140713202	Mother			
		100.749	100.917	167.7	45.3	3	134002005	Proband	Not determined	*EMID2* and *RABL5*	
		100.788	100.917	128.9	38.1	3	134002002	Mother			
		100.749	100.915	166.4	22.8	3	150001004	Proband	Affected sib inherited duplication	*EMID2* and *RABL5*	
		100.749	100.917	167.7	30.9	3	150001002	Mother			
		100.788	100.917	128.9	50.4	3	131085007	Father	Not transmitted to proband or two unaffected sibs	*EMID2*	
		100.788	100.915	127.6	12.7	3	C2142	Control	N/A	*EMID2*	
	2	110.198	110.271	73.0	40.0	1	116325010	Father	Not transmitted to proband	*IMMP2L (intronic)*	No
	3	110.639	111.454	815.6	119.3	3	133023001	Father	Not transmitted to other affected sib or unaffected sib (by QMPSF)	*IMMP2L* and *DOCK4*	Yes (60%)
		110.708	111.450	742.6	56.0	3	133023005	Proband			
	4	115.640	115.791	150.9	10.5	1	UKTS9020	Control	N/A	*TES*	Yes (23%)
	5	124.276	125.280	1003.9	10.3	1	C2048	Control	N/A	*POT1*	Yes (47%)
		124.521	124.856	334.6	12.8	1	130518003	Father	Not transmitted to proband	*No RefSeq genes*	Yes (85%)
	6	132.728	132.843	114.5	32.4	1	130519003	Father	Not transmitted to proband or affected sib	*EXOC4*	No
		132.728	132.829	100.7	11.5	4	C0906	Control	N/A	*EXOC4*	
	7	133.784	133.793	8.7	14.7	1	131117006	Mother	Not transmitted to proband	*AKR1B1*	No

[a] Unless otherwise stated, segregation was determined by haplotype analysis of published genotype data (Szatmari et al 2007)